**Assignment 12.1**

**Problem Statement:**

● Explain the need of Flume.

Apache Flume is a data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events etc. from various sources to a centralized data store.

When one run of product comes off the line, it generates a log file about that run. Even if this occurs hundreds or thousands of times per day, the large volume log file data can stream through Flume into a tool for same-day analysis with Apache Storm or months or years of production runs can be stored in HDFS and analyzed by a quality assurance engineer using Apache Hive.

Flume is used to move the log data generated by application servers into HDFS at a higher speed.

Using Apache Flume we can store the data in to any of the centralized stores (HBase, HDFS).

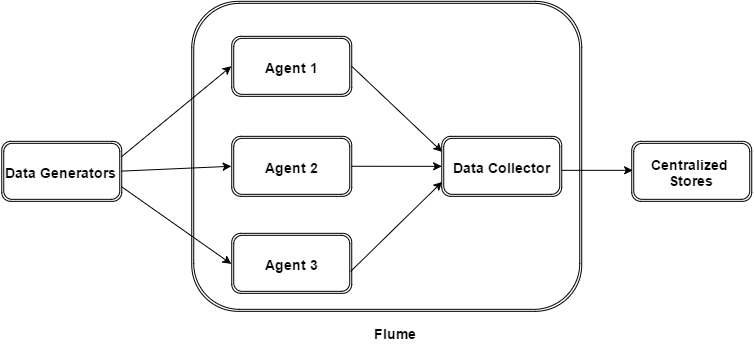
When the rate of incoming data exceeds the rate at which data can be written to the destination, Flume acts as a mediator between data producers and the centralized stores and provides a steady flow of data between them.

Flume provides the feature of contextual routing.

The transactions in Flume are channel-based where two transactions (one sender and one receiver) are maintained for each message. It guarantees reliable message delivery.

Flume is reliable, fault tolerant, scalable, manageable, and customizable.

● Explain the working of Flume and its components in brief.

****

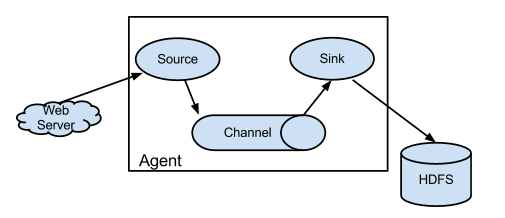
As shown in the illustration,

Flume consists of:

Agents and

Data Collector

data generators (such as Facebook, Twitter) generate data which gets collected by individual Flume agents running on them. Thereafter, a data collector (which is also an agent) collects the data from the agents which is aggregated and pushed into a centralized store such as HDFS or HBase.

****

An agent is an independent daemon process (JVM) in Flume. It receives the data from clients or other agents and forwards it to its next destination. Flume may have more than one agent.

Flume Agent contains three main components namely, source, channel, and sink.

Source

A source is the component of an Agent which receives data from the data generators and transfers it to one or more channels in the form of Flume events.

Apache Flume supports several types of sources and each source receives events from a specified data generator.

Example − Avro source, Thrift source, twitter 1% source etc.

Channel

A channel is a transient store which receives the events from the source and buffers them till they are consumed by sinks. It acts as a bridge between the sources and the sinks.

These channels are fully transactional and they can work with any number of sources and sinks.

Example − JDBC channel, File system channel, Memory channel, etc.

Sink

A sink stores the data into centralized stores like HBase and HDFS. It consumes the data (events) from the channels and delivers it to the destination. The destination of the sink might be another agent or the central stores.